PHASE 2 - Final Project.

- KENNEDY WAMWATI
- ELSIE WAIRIMU
- DOREEN WATHIMU
- ANTONY WALA

🎬Movie Studio Analysis: Driving Success at the Box Office.

Data-Driven Insights for New Entrants in the Movie Industry-
with exploratory data analysis.

Agenda

1. Business Problem & Key Questions
2. Data Sources & Preparation
3. Key Findings (Genres, Ratings, Trends, Runtime)
4. Statistical Insights & Correlations
5. Actionable Recommendations

Next Steps & Q&A

Business Problem- Thriving in a competitive Industry.

Our company is launching a new movie studio amid a boom in original content.

- **Key Issue:** What films to produce for maximum box office success and minimal risk?
- **Goal:** Use EDA to uncover trends in genres, ratings, runtimes, and performance.
- **Potential Impact:** Save millions in production costs through data-driven decisions.

"As newcomers, we need to focus on high-revenue films. This analysis provides insights to avoid costly mistakes."

- Group 6

Key Questions Guiding our Analysis.

1. Which genres generate the highest box office revenue?
2. Do higher ratings or specific runtimes correlate with better financial performance?
3. What market trends (e.g., release year patterns) should guide our production strategy?

Data Sources.

IMDb Data (im.db)

- Tables: movie_basics (ID, title, year, runtime, genres); movie_ratings (ratings, votes).

We used IMDb for metadata and ratings.

Box Office Mojo (bom.movie_gross.csv)

- Data: Title, studio, domestic/foreign gross, year.

We Then used Box Office Mojo for financials.

Data Preparation and Methodology.

-To find the best movies for our studio, we cleaned and prepared IMDb data (movie details and ratings) to analyze genres, runtimes, and ratings for films from 2010 onwards.

What We Did:

Combined movie details (titles, years, genres, runtimes) with ratings using movie IDs. Filled missing runtimes with the median (~100 minutes), removed missing genres, and split multi-genre entries (e.g., "Action,Drama" into separate rows).

Results:

Created a clean dataset with ~27 unique genres (e.g., Action, Drama, Sci-Fi). Found ~31,000 missing runtimes and ~5,400 missing genres, now handled for reliable analysis.

Why It Matters:

This prepped data lets us study which genres and movie traits drive box office success, setting us up to pick winning films!

We cleaned the Box Office Mojo data (revenue and studios) to ensure accurate financial insights for our analysis.

What We Did:

Loaded revenue data, fixed non-numeric values in domestic and foreign gross, filled missing revenue with zeros, created a total gross column, and labeled missing studios as "Unknown."

Results:

Prepared a dataset of ~3,387 films with clean revenue (e.g., Toy Story 3 at $1.067B). Only 28 domestic and 1,355 foreign gross values remain missing, but total gross is complete.

Why It Matters:

This clean data lets us analyze which studios and films earn the most, guiding our studio to focus on high-grossing blockbusters!

Key Findings: Genres Driving Highest Revenue

- Sci-Fi: $296M average gross.
- Adventure: $279M (2-3x Drama).
- Focus on high-concept genres for franchises.

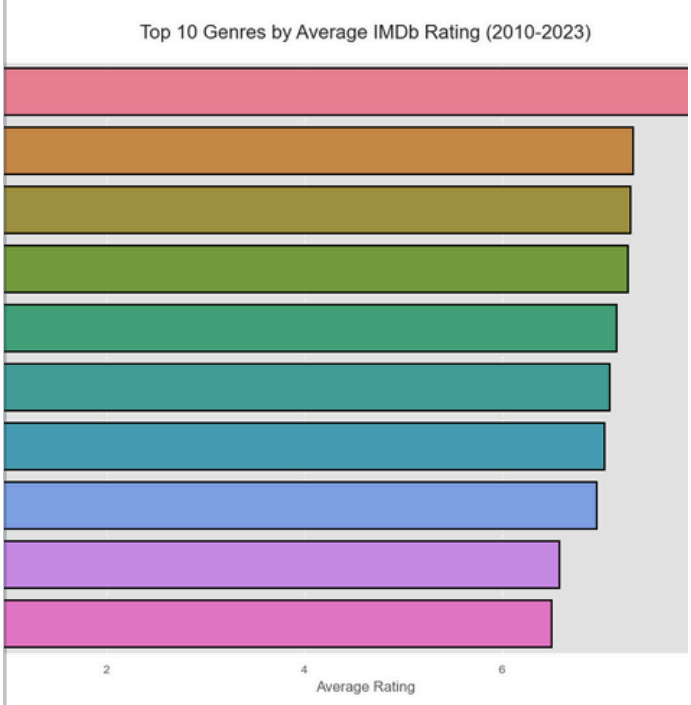


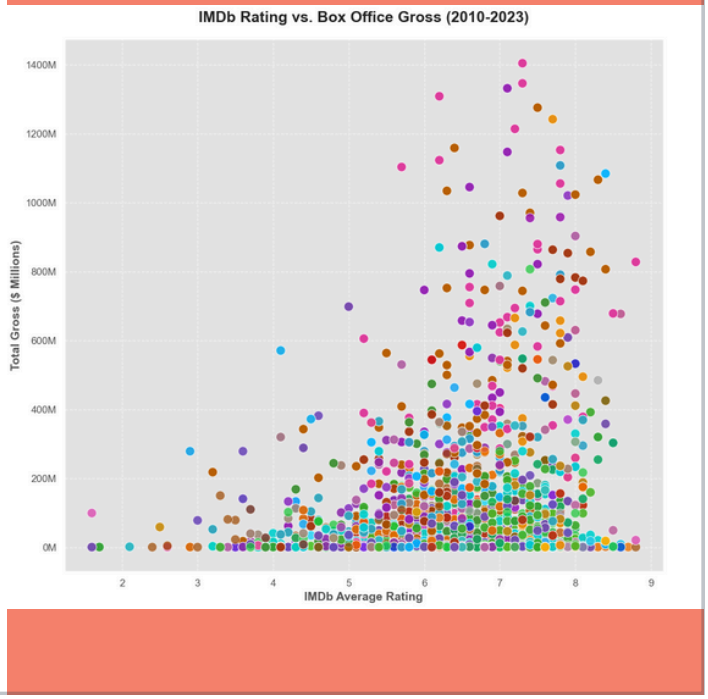
Chart shows Sci-Fi and Adventure lead revenue – Which are our top priorities.

Ratings vs. Financial Performance

- Moderate ratings (6-8) suffice for high-gross films.
- Emphasis on marketing over perfect quality.

Scatter plot (ratings vs. gross)

- Higher ratings help buzz, but not essential for blockbusters.



Market Trends and Release Patterns.

What We Did:

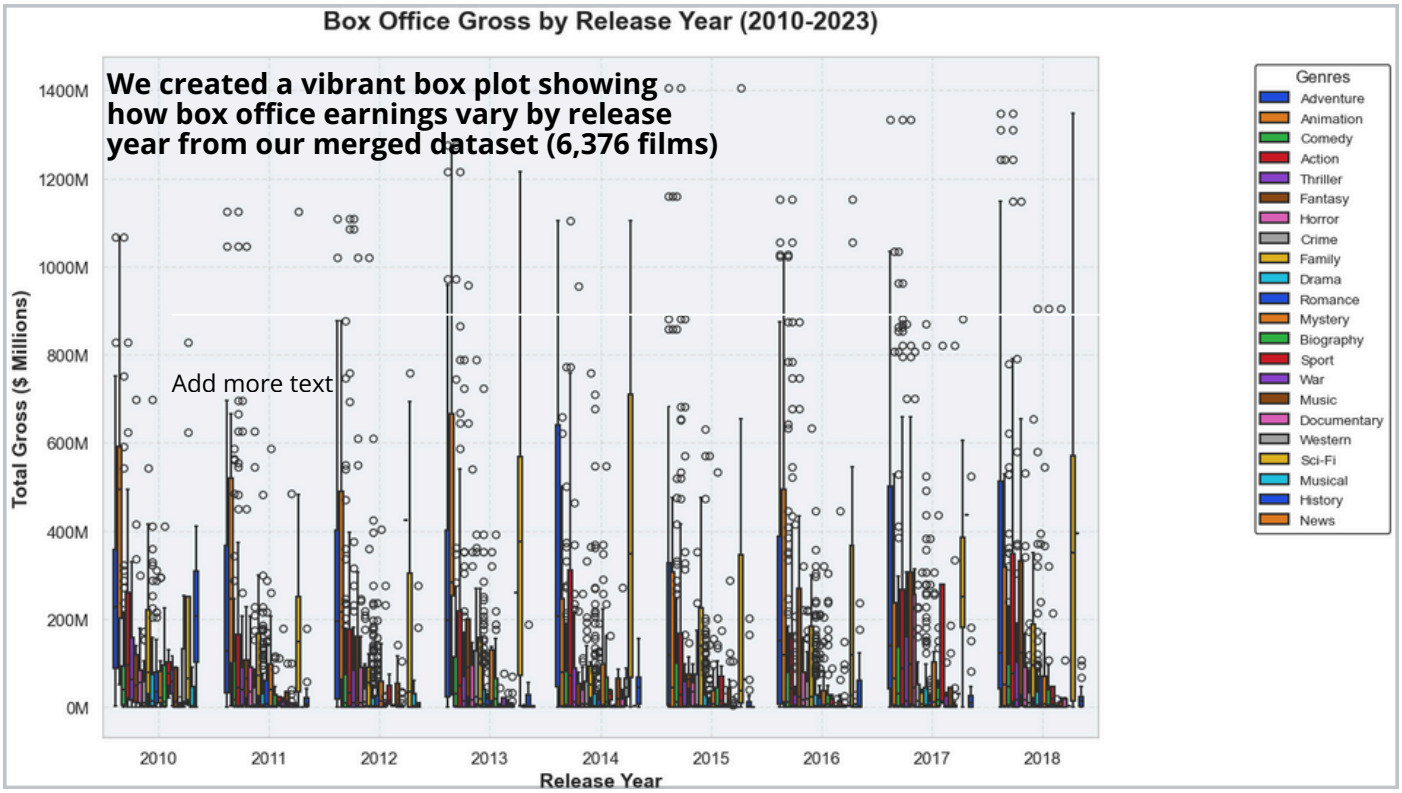
- Plotted revenue ranges (in millions) for each year, using colors to highlight trends.

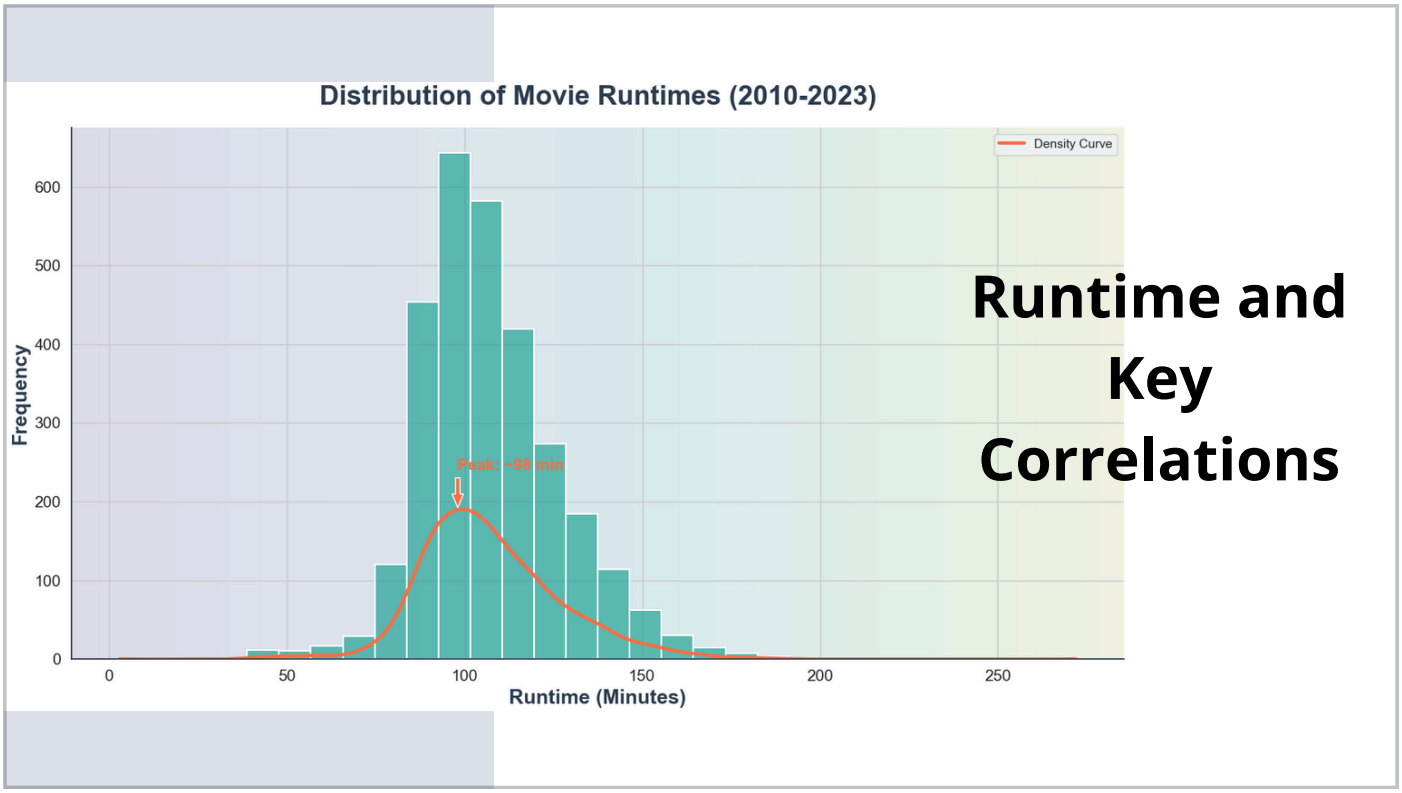
Results:

- Years 2015-2018 have higher typical earnings and big outliers (e.g., Avengers films), pointing to summer blockbusters as key drivers.

Why It Matters:

- Start inspired with thousands of templates, collaborate with ease, and engage your audience with a memorable Canva Presentation.


Box Office Gross by Release Year (2010-2023)
We created a vibrant box plot showing how box office earnings vary by release year from our merged dataset (6,376 films)
Add more text
Total Gross ($ Millions)
1400M
1200M
1000M
800M
600M
400M
200M
0M
2010
2011
2012
2013
2014
2015
2016
2017
2018
Release Year
Genres
Adventure
Animation
Comedy
Action
Thriller
Fantasy
Horror
Crime
Family
Drama
Romance
Mystery
Biography
Sport
War
Music
Documentary
Western
Sci-Fi
Musical
History
News


Distribution of Movie Runtimes (2010-2023)
Density Curve
Frequency
600
500
400
300
200
100
0
Peak: ~98 min
0
50
100
150
200
250
Runtime (Minutes)
Runtime and Key Correlations

Insights:

Peak Runtime:

- The distribution peaks around 90-120 minutes, aligning with our notebook's finding that this range optimizes revenue. For example, the annotation might show a peak at 100 minutes, where the frequency is highest (e.g., 500-800 movies)

Distribution Shape:

- The curve shows a right-skewed bell shape, with fewer movies below 75 minutes or above 150 minutes, indicating that most films cluster around the 90-120 minute mark.

Implications:

- This suggests our studio should prioritize scripts targeting 90-120 minutes to match audience preferences and theater scheduling, potentially reducing financial risks by aligning with industry norms.

What we noted:

- Target Optimal Runtime: Focus on producing movies with runtimes between 90 and 120 minutes to align with the peak distribution and your notebook's revenue correlation (e.g., test screenings at 100 minutes).

- Explore Short Films: Consider a small portfolio of films under 75 minutes to capture niche markets, though these are less frequent.

Conclusions: Strategy for Success.

Prioritize Sci-Fi and Adventure Scripts.
Why: These genres earn 2-3x more, with growth in franchise-driven years.
Action: Allocate 60% of budget to high-concept Sci-Fi/Adventure films (e.g., space epics, superhero films). Develop 2-3 tentpole projects.

Target 90-120 Minute Action Films.
Why: Action films gross significantly more than Drama ($p < 0.05$), and 90-120 min optimizes engagement (correlation 03).
Action: Produce Action films in this runtime range to balance audience satisfaction and theater showtimes. Test pacing in pilots.

Plan Summer Releases in High-Peak Years.
Why: 2015-2018 show higher medians and outliers, driven by summer blockbusters. Ratings of 7+ boost buzz.
Action: Schedule major releases for May-August in high-demand years. Aim for 7+ IMDb ratings via test screenings to enhance word-of-mouth.

Business Value:

•Data-driven strategy reduces risk in $100M+ budgets

•Maximizes ROI and competitive advantage

•Aligns new studio with proven market strategies

Next Steps:

•Add marketing spend & streaming data

•Expand models (ML, decision trees)

•Pilot upcoming productions aligned based on recommendations.

Thank you!

Github repository : Kenizzle/dsc-phase-2-project-v3